DOCUMENT RESUME

ED 450 151 TM 032 341

AUTHOR Jacobs, Robert
TITLE Outliers in Statistical Analysis: Basic Methods of Detection and Accommodation.
PUB DATE 2001-02-00
NOTE 21p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, February 1-3, 2001).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Identification; *Statistical Analysis
IDENTIFIERS *Outliers

ABSTRACT

Researchers are often faced with the prospect of dealing with observations within a given data set that are unexpected in terms of their great distance from the concentration of observations. For their potential to influence the mean disproportionately, thus affecting many statistical analyses, outlying observations require special care on the part of the researcher. It is suggested that decisions about how to go about discarding or incorporating such outliers be made with careful consideration as to the implications associated with the various procedures for doing so. Several methods for dealing with outliers are illustrated. (Contains 2 tables and 10 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 450 151

Outliers in Statistical Analysis:

Basic Methods of Detection and Accommodation

Robert Jacobs

Texas A&M University 77843-4225

R. Jacobs

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM032341

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, February 1-3, 2001.

BEST COPY AVAILABLE

Abstract

Researchers are often faced with the prospect of dealing with observations within a given data set that are unexpected in terms of their great distance from the concentration of observations. For their potential to disproportionately influence the mean, and thus many statistical analyses, outlying observations require special consideration on the part of the researcher. It is suggested that decisions about how to go about discarding or incorporating such outliers be made with careful consideration as to the implications associated with the various procedures for doing so.

In order to review methods of dealing with outliers in a data set, it is first necessary to examine briefly the nature of an outlier in order to make explicit why handling them requires such care. Defining the cutoff for what makes an extreme measurement an outlier requires some subjective judgement. When dealing with a data set that is normally distributed, a particular data point could theoretically exist anywhere within the range of the distribution (Sachs, 1982). As do many authors, Barnett and Lewis (1984) defined an outlier in a set of data as "an observation (or subset of observations) which appears to be inconsistent with that set of data."

The researcher's understanding of the source and degree of this inconsistency may guide the decision on the best method of dealing with an outlier. Usually, the researcher is concerned with the type of variability the outlier represents. Within the broad category of outliers, Beckman and Cook (1983) distinguished between the nature of an outlier as either a discordant observation or as a contaminant. They define a discordant observation as "any observation that appears surprising or discrepant to the investigator" (p. 121). Alternatively, they define a contaminant as "any observation that is not a realization of the target distribution" (p. 121). Clearly, the two types of outliers represent different types of miscalculations on the part of the researcher. It is the latter

type, however, which is potentially more troubling on a technical level.

Without specifically addressing the idea of contaminants, Anscombe (1960) discussed the types of variability that may exist within a sample, including the concepts of measurement error, which is a failure of the measurement instruments, and execution error, which represents a discrepancy between what was intended to be measured and what actually was. When confronted with an observation or set of observations that deviate markedly from the expected range, a thoughtful researcher must consider the accuracy of the measurement, for no observations are absolutely trustworthy (Anscombe, 1960). In certain cases, measurement error may be a result of faulty recording or coding of an observation, or the result of an imperfect measurement apparatus. Indeed, there exist numerous reasons why measurement error may be to blame for an incongruent observation (Anscombe, 1960). In this regard, Iglewicz and Hoaglin (1993) pointed out that identifying the cause of such miscalculations and subsequently amending the procedures whereby they were produced is vastly important in reducing the error rate in future samples.

Sample size also factors into the detection of outliers, as the smaller a sample is, the less probable are outliers (Sachs, 1982). In collecting a sample by means of random selection from

a normally distributed population, the odds favor selecting a data point from where the concentration of data points is highest. As the size of a sample grows, it begins to mirror the population from which it is drawn more accurately (Sachs, 1982).

Evans (1999) discussed the challenge researchers face in determining the cause of outliers. Although the researcher is rarely certain of the singular cause of an outlier, understanding the source of the unexpected variability is important in making a decision on a course of action for dealing with such an observation. Researchers may wish to recode, retain, or eliminate observations altogether.

In addition to measurement and sampling errors, several other sources for outliers exist. For example, outliers can exist as the result of an incorrect distributional assumption (Iglewicz & Hoaglin, 1993). This may be the case if, for instance, the researcher assumes that a sample is normally distributed when in fact it is not. Such a case may exist if the distribution were skewed in one direction, or if it was rectangular. If this is indeed the cause, anomalous observations are not true outliers. Rather, their examination may assist the researcher in adopting a more suitable statistical model that leads to more valid inferences (Iglewicz & Hoaglin, 1993). As Barnett and Lewis (1983) declared, the analysis of the data may sound a warning in terms of the distribution. Douzenis and Rakow

(1987) supported this perspective by pointing out that the presence of an outlier may indicate a weakness in the statistical model and that outliers may distort statistics which assume an interval level of measurement or a normally distributed sample.

Another possible cause of outliers is that the data contains a different structure than is accounted for by the sampling method. Iglewicz and Hoaglin (1993) used the example of a researcher employing random daily samples when in fact morning and afternoon subsamples might be more appropriate and accurate. In cases such as this, qualitative differences between data points may exist within a sample that could produce the appearance of outliers. Barnett and Lewis (1983) used an example of data that refer to purchases of cereal packets over a 13 week period, where 2 apparent outliers are recorded at 52 and 39 packets sold. It is suggested that purchasing trends unseen by the researchers may account for these outliers, as they are both multiples of 13.

Another factor to consider in studying outliers is that they may not be mistakes at all, but indicators that within a sample, such values are possible (Iglewicz & Hoaglin, 1993). In such a case, the presence of an outlier can lead the researcher to an important discovery in terms of the potential for what is being studied.

While examining a set of data points, researchers face the difficulty of accurately identifying outliers. Visual inspection alone, without the use of analytic or graphical tools, is an inadequate means of analysis in this regard, and can lead to missing or mislabeling outliers (Iglewicz & Hoaglin, 1993).

One method of identifying outliers would simply be to convert the data points to Z scores and screen for high absolute values (Donzenis & Rakow, 1987). A Z score is the observed value minus the mean, divided by the standard deviation. A data point's Z score represents the number of standard deviations it falls from the mean. Donzenis and Rakow (1987) suggested that Z scores of plus or minus 2.70 should be considered "outside" because they are "1.5 times the interquartile range (a step) below the twenty-fifth or above the seventy-fifth percentiles" (p. 4). In turn, Donzenis and Rakow (1987) suggested that Z scores of plus or minus 4.72 should be considered "far out", because those values are "beyond three times the interquartile range (two steps) below the twenty-fifth or above the seventy-fifth percentiles" (p. 4.).

Although this method of identifying outliers is appealing for it's simplicity, Iglewicz and Hoaglin (1993) pointed out the method's inherent inaccuracy. The maximum possible Z score for each data point is constrained by the subtraction of the mean and the division of the standard deviation. A large difference

between the data point and the mean contributes to a large standard deviation, which confines the resulting Z score.

Consider this small heuristic data set: 1.0 .91 1.04 .89 1.20 .90 1.10 2.0.

For these data points, the mean is equal to 1.13. The standard deviation is then .361. The respective Z scores then are: .360 .61 .249 .665 .194 .637 .08 2.37.

Given this method, a researcher would not conclude that the score of 2.0 with Z = 2.37 is an outlier, though it is clearly a departure from the other data points. The Z score of 2.37 falls short of the proposed criteria for categorizing a data point as "outside", let alone the criteria for what constitutes consideration for "far out". The process of identifying outliers by their Z scores is particularly problematic in small data sets, because as Iglewicz and Hoaglin (1993) pointed out, the maximum Z score for any given data point is equal to (n-1) divided by the square root of n.

The Box Plot

One method of identifying outliers graphically is by examining a box plot (Tukey, 1977). A box plot's measure of central tendency is the median, beneath which 50% of the data points fall. The "box" surrounding the median represents the 75^{th} and 25^{th} percentiles of the data. This is known as the interquartile range. Additionally, the highest and lowest scores

are represented. Tukey (1977) suggested that outliers can be determined by the use of "fences" around the interquartile box. The fences may be drawn to extend 1.5 to 3 times the difference between the first and third quartiles, depending on the researcher's interest in sensitivity in determining an outlier. Thus, observations that fall beyond these fences are viewed by this method as outliers. Figure 1 shows a box plot of the heuristic data set used earlier. As the difference between the first and third quartiles is equal to .25, drawing the fence to 1.5 times the difference would extend the boundary for an outlier to .375 above and below the interquartile range. Using a more conservative multiple of 3.0, the fence would extend .81 beyond the interquartile range. Within this data set, either method would mark a score of 2.0 as a clear outlier. Most striking, however, is the box plot's utility in making outlying scores of stand out against the concentration of data points.

INSERT FIGURE 1 ABOUT HERE.

Statistical Significance Testing

Sachs (1982) suggested several methods for testing a data set for outliers based on the size of a random, normally distributed sample. The first is to be used for relatively small samples, where n is less than or equal to 25. First, the individual values are ranked in terms of magnitude. The

researcher then finds the difference between the suspected outlier and the observation closest to it, then divides this value by the difference between the suspected outlier and the value of the least magnitude. As the data set increases in size to where n is equal to between 8 and 13 the equation changes slightly so that denominator is computed by finding the difference between the suspected outlier and 1 minus the observation of the least magnitude. When the data set is between 14 and 25, the denominator is computed by finding the difference between the suspected outlier and two less than the observation of the least magnitude. The result of this equation is then compared with specified significance values. see Table 1.

Consider the heuristic data set of: 1.0 .91 1.04 .89 1.20 .90 2.0. In order of magnitude, the data set is: .89 .90 .91 1.0 1.04 .120 2.0.

Using the formula in Sachs (1982), where 2.0 is the suspected outlier, we find that: 2 - 1.20/ 2 - .89 is equal to .870. The critical value in Table 1 for n=7 at a .05 statistical significance level is .507. Because .870 exceeds this values, the null hypothesis that there are no outliers in this data set is rejected.

INSERT TABLE 1 ABOUT HERE

For data sets larger than 25, Sachs (1982) provided a test statistic and a table of critical values. If the value of the test statistic, T, equals or exceeds the upper statistical significance boundary of the standardized extreme deviation, it can be assumed that the extreme value originated from a population other than the data points in the remainder of the sequence (Sachs, 1982). An approximately normally distributed sample is assumed for this test. Table 2 provides the critical values.

For this test, the formula is:

$$T = x1 - \bar{x} / s$$

Here, x1 is the extreme value and s is the standard deviation of the sample.

For instance, consider a sample of 50, where the mean was equal to 100 and the standard deviation was equal to 15. If the extreme value was recorded at 175, the formula would be:

$$T = 175 - 100 / 15.$$

This formula results in the value 5.0. From Table 2, we see that the critical value for a sample of this size, at a 95% statistical significance level is equal to 3.083. Thus, by exceeding this critical value, it could be assumed that the

extreme score is representative of a different population than the remainder of sample represents.

INSERT TABLE 2 ABOUT HERE

This paper excludes multiple methods of detecting outliers statistically for the purpose of regression analysis. The interested reader can find an in Evans (1999) an exceptional primer on this topic in particular.

Accommodating Outliers

Once an outlier has been identified, the researcher is left with the question of how to deal with it, for it's presence impacts statistical analysis, perhaps most basically by affecting the mean. As is widely recognized, the mean is disproportionaly impacted by outliers (Wilcox, 1997, 1998). Because other statistics employ deviations from the mean, distortions in the mean in turn distort other statistics.

It is best to conduct statistical analyses both including and excluding the potential outliers once they have been identified. The difference between the two analyses can guide the researcher in the best course of action for dealing with the outlier (Sachs, 1982).

One option is simply to discard the outlier. Sachs (1982) suggested that in a data set of at least 10 individual values, a value may be discarded if it exceeds 4 standard deviations from

the mean, where the mean and standard deviation are computed without the suspected outlier. Sachs (1982) points out that 99.9% of the data exists within 4 standard deviations from the mean in a normal distribution and that 97% of the data exists within this range in symmetric, unimodal distributions.

Another method is to use a "trimmed" mean. As the mean is particularly vulnerable to being influenced by outlying observations, computing a trimmed mean allows for a potentially more reliable estimator of central tendency (Iglewicz & Hoaglin, 1993). The process of trimming the mean is not necessarily one designed to accommodate outliers, though it can certainly be used for that purpose (Sachs, 1982). To trim the mean, a researcher eliminates the observation or observations that have been flagged as potential outliers. To retain the symmetry of the sample the researcher then eliminates an observation or observations at the opposite end of the sample. For accuracy purposes it is best not to trim more than 15% of the sample when computing the mean (Iglewicz & Hoaglin, 1993). Thus, the mean is computed using the trimmed sample. However, the eliminated data points are then reinserted into the sample for the remaining analyses.

It is simplest to illustrate the process of trimming the mean with a small data set of, say, 11 observations. Consider this heuristic data set: 100 97 91 109 116 89 101 119 87

92 191. To compute the trimmed mean, the data points are first ordered in terms of magnitude: 87 89 91 92 97 100 101 109 116 119 191.

After having identified the highest score of 191 as a potential outlier, it is dropped temporarily from the data set, as is the lowest score of 87. The original sample yields a mean of 108.36. The trimmed mean is 101.56, which resembles more clearly the whole of the sample.

Another method of accommodating outliers is to compute a winsorized mean. Winsorization is an alternative to trimming which substitutes a copy of the adjacent value in a data set for the observation that is considered an outlier (Sachs, 1982). Iglewicz and Hoaglin (1993) proposed the process of winsorization be parallel to the mechanics of trimming, that is, that trimming should be performed on a data set symmetrically, with an equal number of values trimmed from each end of the data set. In this case, to compute a winsorized mean, the researcher replaces observations of both high and low magnitude with the adjacent observations in the data set.

To illustrate the process of winsorization, consider the heuristic data set from the previous example: 100 97 91 109 116 89 101 119 87 92 191. The researcher first orders the data points in terms of their magnitude: 87 89 91 92 97 100 101 109 116 119 191.

As with the trimmed mean, the scores of 191 and 87 are dropped from the data set. Next, the scores closest to these values in terms of magnitude replace the dropped scores in the data set. The winsorized mean is the computed as:

89 + 89 + 91 + 92 + 97 + 100 + 101 + 109 + 116 + 119 + 119/11.

Thus, the winsorized mean is equal to 102.0. Recall that the trimmed mean is equal to 101.56, and that the mean of the uncorrected data set equals 108.36. This slight increase over the magnitude of the trimmed mean illustrates that computing a winsorized mean does not discard outliers completely, but rather decreases their distance from the center of the sample (Barnett & Lewis, 1984). The interested reader can find a more complete treatment of the multiple facets and uses for trimmed and winsorized means both in Barnett and Lewis (1984) and in Beckman and Cook (1983).

References

Anscombe, F.J. (1960). Rejection of outliers. Technometrics, 2, 123-147.

Barnett, V. & Lewis, T. (1984). Outliers in statistical data. Chichester, England: John Wiley & Sons.

Beckman, R.J. & Cook, R.D. (1983). Outlier....s. Technometrics, 25, 119-149.

Douzenis, C. & Rakow, E.A. (1987, November). Outliers: a potential data problem. Paper presented at the annual meeting of the Mid-South Educational Research Association, Mobile, AL. (ERIC Document Reproduction Service no. ED 291 789)

Evans, V. (1999, January). Strategies for detecting outliers in regression analysis: An introductory primer. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX. (ERIC Document Reproduction Service no. ED 427 059)

Iglewicz, B. & Hoaglin, D.C. (1993). How to detect and handle outliers. Milwaukee, WI: ASQC Quality Press.

Sachs, L. (1982). Applied statistics: A handbook of techniques (2nd ed.). New York: Springer-Verlag.

Tukey, J.W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

Wilcox, R.R. (1997). Introduction to robust estimation and hypothesis testing. San Diego: Academic Press.

Wilcox, R.R. (1998). How many discoveries have been lost by ignoring modern statistical methods? American Psychologist, 53, 300-314.

Figure 1

Boxplot of values from the heuristic data set

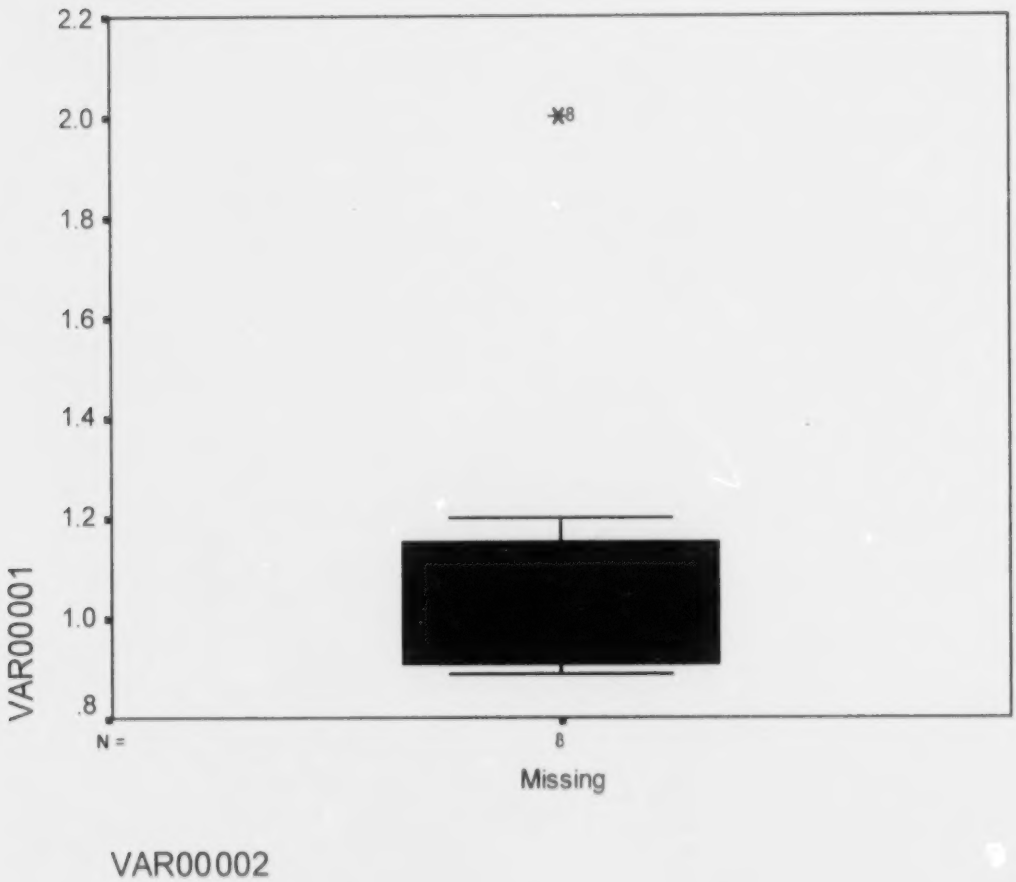

Table 1
Significance Bounds for Testing a Smaller Sample for the Presence of an Outlier

N	alpha=.10	alpha=.05	alpha=.01	Test statistic
3	0.886	0.941	0.988	
4	0.679	0.76	0.889	
5	0.557	0.642	0.780	
6	0.482	0.560	0.698	x1 - x2
7	0.434	0.507	0.637	x1 - xn
8	0.479	0.554	0.683	
9	0.441	0.512	0.635	x1 - x2
10	0.409	0.477	0.597	x1 - xn-1
11	0.517	0.576	0.679	
12	0.490	0.546	0.642	
13	0.467	0.521	0.615	
14	0.492	0.546	0.641	
15	0.472	0.525	0.616	
16	0.454	0.507	0.595	
17	0.438	0.490	0.577	
18	0.424	0.475	0.561	
19	0.412	0.462	0.547	
20	0.401	0.450	0.535	
21	0.391	0.440	0.524	
22	0.382	0.430	0.514	
23	0.374	0.421	0.505	
24	0.367	0.413	0.497	x1 - x2
25	0.360	0.406	0.489	x1 - x n-2

Table 2
Significance Bounds for Testing a Larger Sample for the Presence of an Outlier

N	S=95%	S=99%	N	S=95%	S=99%
1	1.645	2.326	55	3.111	3.564
2	1.955	2.575	60	3.137	3.587
3	2.121	2.575	65	3.160	3.607
4	2.234	2.806	70	3.182	3.627
5	2.391	2.877	80	3.220	3.661
6	2.386	2.934	90	3.254	3.691
8	2.490	3.022	100	3.283	3.718
10	2.568	3.089	200	3.474	3.889
15	2.705	3.207	300	3.581	3.987
20	2.799	3.289	400	3.656	4.054
25	2.870	3.351	500	3.713	4.106
30	2.928	3.402	600	3.758	4.148
35	2.975	3.444	700	3.797	4.183
40	3.016	3.479	800	3.830	4.214
45	3.051	3.511	900	3.859	4.240
50	3.083	3.539	1000	3.884	4.264